The role of visual spatial attention in audiovisual speech perception

Andersen, Tobias; Tiippana, K.; Laarni, J.; Kojo, I.; Sams, M.

Published in:
Speech Communication

Link to article, DOI:
10.1016/j.specom.2008.07.004

Publication date:
2009

Document Version
Early version, also known as pre-print

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2008) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The role of visual spatial attention in audiovisual speech perception ☆

Tobias S. Andersen [a,b,*], Kaisa Tiippana [c], Jari Laarni [d], Ilpo Kojo [e], Mikko Sams [c]

[a] Center for Computational Cognitive Modeling, Department of Psychology, University of Copenhagen, Linnésgade 22, 1361 Copenhagen K, Denmark
[b] Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 321, 2800 Lyngby, Denmark
[c] Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, P.O. Box 9203, 02015 TKK, Finland
[d] VTT Technical Research Centre, P.O. Box 1000, 02044 VTT, Finland
[e] Center for Knowledge and Innovation Research, Helsinki School of Economics, P.O. Box 1210, 00101 Helsinki, Finland

Received 18 April 2007; received in revised form 23 July 2008; accepted 23 July 2008

Abstract

Auditory and visual information is integrated when perceiving speech, as evidenced by the McGurk effect in which viewing an incongruent talking face categorically alters auditory speech perception. Audiovisual integration in speech perception has long been considered automatic and pre-attentive but recent reports have challenged this view. Here we study the effect of visual spatial attention on the McGurk effect. By presenting a movie of two faces symmetrically displaced to each side of a central fixation point and dubbed with a single auditory speech track, we were able to discern the influences from each of the faces and from the voice on the auditory speech percept. We found that directing visual spatial attention towards a face increased the influence of that face on auditory perception. However, the influence of the voice on auditory perception did not change suggesting that audiovisual integration did not change. Visual spatial attention was also able to select between the faces when lip reading. This suggests that visual spatial attention acts at the level of visual speech perception prior to audiovisual integration and that the effect propagates through audiovisual integration to influence auditory perception.

Keywords: Speech perception; Multisensory; Attention; McGurk effect

1. Introduction

The audiovisual nature of speech perception is demonstrated by two perceptual phenomena. First, watching congruent articulations enhances speech perception (Sumby and Pollack, 1954). This enhancement is particularly strong in noisy environments (Sumby and Pollack, 1954) and for the hard of hearing (Grant and Seitz, 1998) – i.e. when auditory speech is less reliable. Second, in the so-called McGurk effect, watching an incongruent articulation categorically alters the auditory speech percept (McGurk and MacDonald, 1976). The classical example of this effect is that the auditory speech token /ba/ is perceived as /da/ when watching a face saying /ga/ (McGurk and MacDonald, 1976).

From the first report, most studies have considered the McGurk effect to happen pre-attentively, or automatically. Two reasons were given in the first report by McGurk and MacDonald (1976). First, knowledge about the illusory and incongruent nature of the audiovisual speech stimuli does not disrupt the illusion. This has been confirmed by informal reports (Liberman and Mattingly, 1985). Second, extended exposure to McGurk type stimuli does not decrease the effect. Considering the crucial role of attention in binding features into complex objects suggested by Treisman and Gelade (1980), it would be surprising if an object as complex as a talking face should be perceived pre-attentively. It is of course possible that audiovisual feature

☆ Preliminary results from this study was presented at the International Multisensory Research Forum, Geneva, Switzerland, 2002.

* Corresponding author. Address: Center for Computational Cognitive Modeling, Department of Psychology, University of Copenhagen, Linnésgade 22, 1361 Copenhagen K, Denmark. Tel.: +45 4525 3687; fax: +45 4588 2673.
E-mail address: ta@imm.dtu.dk (T.S. Andersen).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.07.004

integration is governed by different mechanisms than intramodal feature integration. It is also possible that speech perception is a special mode of perception, which is not dependent on attention (Liberman and Mattingly, 1985, 1989). The question of whether cross-modal feature integration in speech perception requires attention thus addresses fundamental theoretical topics. Furthermore, since face-to-face conversation is the most important mode of communication for humans, and speech-specific disorders only too common, understanding the role of attention in audiovisual speech perception may also have clinical impact.

In addition to McGurk and McDonald's report work by Massaro (1987, 1998) has been influential in conveying the view that audiovisual speech perception occurs pre-attentively. Massaro studied the effect of intermodal attention on audiovisual speech perception. He asked participants to report either what they heard or what the saw and found a significant difference between the two conditions. Since this manipulation did not completely disrupt cross-modal effects, Massaro concluded that audiovisual speech perception is robust to manipulations of attention through task instructions. We find that the conclusion might as well be that attention did indeed have an effect. Massaro also fitted the Fuzzy Logical Model of Perception (FLMP) to the data and found no significant difference in the goodness-of-fits between the two conditions which he took as evidence that audiovisual integration occurred according to the same rule (i.e. the FLMP) in both conditions and was thus unaffected by manipulations of instructions. However, if the FLMP is not a correct and complete model of audiovisual integration of speech – a criticism raised by several studies (Andersen et al., 2002; Crowther et al., 1995; Cutting et al., 1992; Grant and Seitz, 1998; Pitt, 1995; Schwartz, 2006; Vroomen and de Gelder, 2000) – then it cannot be used to determine the effect of attention on audiovisual integration. Massaro (1984) studied the effect of visual attention on audiovisual speech perception, with children as participants. He found that having participants perform an additional task of detecting mouth movements during an audiovisual speech perception task had no effect on performance. If this additional task increased attention towards the mouth region of the face, this result points to no effect of visual attention. However, Tiippana et al. (2004) attributed this result to a ceiling effect so that, since objects at fixation tend to be attended, in Massaro's design there was little room for improvement with enhanced attention.

Soto-Faraco et al. (2004) also found support for the audiovisual integration of speech being pre-attentive in a study of a syllabic interference task using audiovisual speech stimuli. In the syllabic interference task, perception of the first syllable of a disyllabic nonsense word is influenced by the second syllable if the second syllable varies from trial to trial but not if it is constant. The second syllable interfering with perception of the first syllable is interpreted as auditory attention failing to select the relevant stimulus so that the second syllable is obligatorily processed even though observers attempt to focus their attention only on the first syllable. In their study, Soto-Faraco et al. used audiovisual stimuli where the talking face would sometimes be incongruent with the second syllable thus creating a McGurk illusion. They found that it was the illusory percept rather than the actually presented acoustic stimulus that determined whether syllabic interference occurred. Soto-Faraco et al. concluded that the visual influence on the auditory speech percept must have occurred before auditory attentional selection.

Some neurophysiological studies have shown that the mismatch negativity (MMN) arises during audiovisual speech perception for changes in visual speech (Colin et al., 2002; Möttönen et al., 2002; Sams et al., 1991). In the MMN paradigm a sequence of sounds is presented. The sequence consists of two types of sounds: standard sounds which occur frequently, and odd sounds, which occur rarely. The MMN is the difference in the neuroelectrical/electromagnetic fields elicited by standards and odds. The MMN is elicited even when observers do not pay attention to the sound sequence, and the MMN is generally believed to reflect a pre-attentive mechanism that alerts the organism to a change in the acoustic environment (Näätänen, 1992). In the above-mentioned audiovisual speech studies using the MMN paradigm, the face and voice matched in the standard stimulus. In the odd stimulus, the voice was the same as in the standard stimulus, but the face articulated a different syllable thus creating a McGurk illusion. Despite the acoustic parts of the standard and odd stimulus being identical, the MMN was elicited. This shows that vision influenced activity in the auditory system at the level of the MMN. If the MMN is considered to be pre-attentive, then these results can be interpreted to suggest that visual influence on audition is too.

Several studies have thus come to the conclusion that auditory attention does not influence the McGurk effect and audiovisual integration of speech. But attention is a complex concept and we cannot a priori extrapolate the effects described above to all faculties of attention. This is emphasized by three recent reports that indicate that there might well be attentional mechanisms involved in audiovisual speech perception.

Tuomainen et al. (2005) presented strong evidence that observers' anticipation of the speech-like nature of sine-wave speech (SWS) has a strong effect on audiovisual integration. They created sine-wave speech by placing a sine-wave tone at the center of each of the three lowest formants of a natural speech signal. When naïve observers are presented with sine-wave speech, they most often do not recognize it as speech, but once they are instructed about the speech-like nature of the stimulus, they are able to understand it (Remez et al., 1981). Tuomainen et al. first trained naïve participants to categorize sine-wave speech tokens in arbitrary categories and found no significant effect of concurrently presented visual speech. But, after instructing the participants of the speech-like nature of

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

the stimulus, they found a strong McGurk effect. Since the stimulus was the same in the two conditions, this shows that audiovisual integration of speech is not entirely a stimulus-driven process and thus that it is influenced by cognitive factors. Tuomainen et al. suggested that observers enter a speech-specific mode of perception only when aware of the speech-like nature of SWS. They interpreted this speech mode in the context of attention by suggesting that it consist of an attentional focus on acoustic features relevant for phonetic classification. These features find counterparts in visual speech and are therefore influenced by it. Contrary, when unaware of the speech-like nature of SWS, observers do not enter speech mode and may focus on acoustic features irrelevant for phonetic classification. These features find no counterpart in visual speech and are therefore unaffected by it.

Alsius et al. (2005) showed that high cognitive load can diminish audiovisual integration of speech. Cognitive load refers to load on working memory and thus on executive systems that control attention. This concept stands in contrast to perceptual load which is a load on early stages of sensory systems due to scene complexity (Lavie, 2005). Alsius et al. increased cognitive load by imposing a secondary object identification task on a primary speech identification task. They found a decrease in the strength of the McGurk effect when participants performed the secondary task. This decrease was not accompanied by a decrease in unimodal speech comprehension which indicates that the effect is likely to occur at the level of audiovisual integration rather than at the level of unimodal perception. This was further supported by the decrease occurring regardless of whether the secondary task was performed on auditory or visual stimuli. Had the effect occurred at the level of unimodal perception, one would expect that an auditory object identification task would interfere with auditory perception to decrease the relative influence of audition and thus *increase* the strength of the McGurk effect.

Tiippana et al. (2004) studied the effect of visual attention on audiovisual speech perception. In the critical condition, attention was distracted by instructing participants to ignore the face and to attend to an image of a moving leaf that was superimposed on the face. The leaf was near the mouth during articulation, so that the effect of eccentric viewing of the face could be presumed to be negligible. In the control condition, participants attended the face and reported what they heard the talker say. Tiippana et al. found that when the leaf was attended the amount of visual influence on the auditory percept decreased showing that visual attention had an effect on audiovisual speech perception.

These three studies all indicate that attention can influence audiovisual speech perception although the aspects of attention they investigated are quite different. Where Tuomainen et al. investigated the effect of speech mode, Alsius et al. addressed the effect of cognitive load and Tiippana et al. addressed the effect of distracting visual object-based attention. In the current study, we will examine the effects of visual spatial attention on audiovisual speech perception using a novel experimental paradigm.

As Tiippana et al. noted, the effect of visual attention on audiovisual speech perception could occur at the level of audiovisual integration of speech. It could also occur at the level of unimodal visual speech perception, i.e. lipreading, and then propagate through audiovisual integration and thus influence audiovisual speech perception. Tiippana et al. discussed this issue at length and found that they were not able to discriminate between these two effects in their study. The current study aims to determine whether effects at these levels can be separated.

Attention as it was manipulated in Tiippana et al.'s study can be divided into two components. One component was *the object* of visual attention. In the critical condition, the object of attention was the leaf, which was unrelated to speech perception and would not cause the McGurk effect to occur. In the control condition, the object of attention was the talking face, which did cause the McGurk effect to occur. Thus, changing the object of attention, from face to leaf, could have decreased the amount of visual influence on speech perception. The other component of attention was *the load*. In the critical condition, the cognitive load of gaze tracking the moving leaf was arguably higher than in the control condition where participants' gaze was fixed on the stable talking face. If cognitive load influences audiovisual speech perception, as Alsius et al. found, this difference could also have caused the decrease of visual influence on speech perception in the critical condition. The current study aims to study the effect of the object of visual attention in the absence of a difference in the cognitive load.

To these aims, we designed stimuli so that the influence of the target face, the visual distractor and the voice could be discerned in participants' responses. The critical, bilateral stimulus consisted of a movie of two talking faces dubbed with a synchronized voice. The faces were displaced symmetrically to the sides of a central fixation point and a cueing arrow pointed to the target face. This stimulus configuration is adapted from Posner's studies of endogenous visual spatial attention (Posner, 1980; Posner et al., 1980). The stimulus configuration is depicted in Fig. 1 (see online Supplementary material for a sample video).

Fig. 1. A sample frame from a bilateral stimulus.

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

One face was saying /aka/, the other face /ata/ while the voice always said /apa/. The stimuli were chosen so that visual /ata/ with auditory /apa/ created the McGurk effect of hearing /ata/; visual /aka/ with auditory /apa/ created the McGurk effect of hearing /aka/ while auditory /apa/ by itself was most often heard veridically as /apa/. Thus, it was possible to discern the relative influence of the target face, distractor face and voice on participants' responses.

If the object of visual spatial attention influences audiovisual speech perception in this paradigm, we will see a striking effect: As *visual* attention is directed from one face to the other, the *auditory* speech percept will change categorically. And, this cannot be due to a difference in cognitive or perceptual load (Lavie, 2005), which is the same regardless of which face is attended. In addition, *how* perception changes is determined by the stage of perception at which attention acts. First, if the auditory speech percept changes towards the utterance of the attended face, so that participants give more T-responses when attending the face saying /ata/ and more K-responses when attending the face saying /aka/, then the effect is one of changing the relative influence of the visual objects – i.e. the faces. This effect could occur already at the level of visual perception and then propagate to audiovisual perception. To test whether this effect occurred at the level of visual perception, we also employed unimodal visual stimuli where no voice was dubbed onto the movie and the participants were instructed to lipread. Second, if the auditory speech percept changes towards the utterance of the voice, so that attending one face rather than the other gives more P-responses then the effect would be one of changing the relative influence of audition and vision. This would indicate that one face influenced auditory perception less than the other. To test this, we employed unilateral audiovisual stimuli in which there was only one face dubbed with the voice. If one face influences speech perception less than the other we would expect more P-responses both when it is presented unilaterally as well as when it is attended in the bilateral stimulus.

In bilateral trials, when the stimulus contained two faces, the perceptual load (Lavie, 2005) is somewhat higher than in unilateral trials where there is only one face. If the level of perceptual load influences audiovisual speech perception, we expect to see a difference between unilateral and bilateral trials. Again, if there is a difference in the proportion of K- and T-responses, this would point to an effect at the level of visual perception. This would indicate that attention occasionally or partly lapsed towards the distractor face. In that case, we expect to see the same difference between unilateral and bilateral trials when there is no voice and participants lipread. Another possibility is that the greater perceptual load in bilateral trials would cause attention to lapse towards the voice which would result in more P-responses. This would indicate that the perceptual load influenced the relative influence of audition and vision and thus affected perception at the level of audiovisual integration.

Finally, we address the issue of eye movements. Although Tiippana et al. argued that effects of eye movements were negligible in their study they did not monitor eye movements directly. In the current study, we monitored participants' eye movements using an eye tracker to ensure correct fixation.

2. Methods

2.1. Stimuli

The auditory stimuli were recorded utterances of /aka/, /apa/ or /ata/ presented through stereo loudspeakers balanced so that the sound appeared to come from the center of the screen. The sound level was 54 dB(A) with a constant background noise level of 43 dB(A). The unilateral visual stimuli were videos of a face saying either /aka/ or /ata/ displaced either to the right or left from a central fixation cross. A cueing arrow, just below the fixation cross, pointed to the face. The viewing distance was 80 cm. The distance from fixation to the center of the mouth was 7.1 cm (5° visual angle). The height of the frame containing the face was 10.9 cm. The face appeared concurrently with the cueing arrow and the fixation point. Mouth opening commenced 2.1 s thereafter. The bilateral visual stimuli contained two faces symmetrically displaced to the sides of the central fixation cross. Fig. 1 displays a sample frame from a bilateral stimulus (see online Supplementary material for the full video). One face said /aka/ while the other said /ata/. The dimensions of the faces and the temporal sequence of the movies were the same as for the unilateral stimuli. The cueing arrow just below the fixation cross indicated which face to attend. The visual stimuli were thus characterized by three factors: Attended visual articulation (V), which could be /aka/ or /ata/, unilateral versus bilateral stimulus (UB) and whether attention was directed to the left or to the right (LR). Audiovisual stimuli consisted of the visual stimuli dubbed with the speaker's voice saying /apa/. The dubbed voice was synchronized with the original voice by aligning the plosive bursts. For bilateral stimuli the plosive burst of the original voices was within one video frame of 40 ms and the dubbed voice was aligned with the best compromise between the two. Audiovisual asynchrony was thus approximately 40 ms which should have little effect on audiovisual integration (Massaro et al., 1996; Munhall et al., 1996).

2.2. Participants

Seven female and seven male (mean age 20.2 years) observers participated in the experiment. All were native speakers of Finnish and naïve as to the purpose of the experiment. The participants reported normal hearing and showed a possibly corrected visual acuity at 80 cm of at least 0.8 decimal.

Pilot studies showed that our visual /ata/ exerted a weak influence on audition. Therefore, in order to make the

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

visual influence on perception stronger, we used a low auditory signal-to-noise ratio. At the same time, auditory speech tokens should be identifiable by themselves. As we did not adjust the auditory signal-to-noise ratio to individual hearing thresholds, two participants never perceived auditory /apa/ correctly, and were therefore excluded from further analysis. In another two participants, the visual face stimuli had no effect on auditory perception in that they perceived /apa/ veridically in 95% of all audiovisual trials where all other participants experienced the McGurk illusion frequently. These two participants were also excluded from further analysis. Thus 10 of 14 participants were included in the following analysis.

2.3. Procedure

Written instructions specified that the participants should maintain their gaze at the fixation point while covertly attending the face indicated by the cueing arrow. The instructions emphasized that they should respond according to what they heard in auditory and audiovisual trials. In visual trials, they were instructed to lipread. Participants could respond with any one or two consonants. A few sample trials were performed to ensure that the instructions were understood.

Each unimodal visual and audiovisual stimulus was presented 12 times evenly distributed between six consecutive blocks. The occurrence of unilateral and bilateral stimuli (UB), and attended visual articulation (V) varied pseudo-randomly within blocks. Attended side (LR) alternated between blocks because our pilot studies indicated that the effect of attention was significantly smaller if attention was to be shifted from side to side between trials possibly due to increased task difficulty or the cost imposed by shifting attention as suggested by Posner (1980). Our initial analysis found no effect of laterality (LR) and participants' responses were therefore pooled across this factor to yield 24 responses per observer at each level of the remaining factors UB and V. In a final block, unimodal auditory /apa/ was presented 24 times, while auditory /aka/ and /ata/ were presented 12 times. The order of presentation of stimuli within a block varied pseudo-randomly.

Double responses with two identical consonants were classified as the corresponding single consonant response. Responses of voiced consonants (G, B and D) were categorized as their unvoiced counterparts (K, P and T). After these re-classifications, responses other than K, P and T were classified as "other".

The participants' eye movements were recorded using a head-mounted gaze tracking system (SMI iView). The right eye was monitored with a miniature infra-red camera while one infra-red LED illuminated the eye. Video images of the pupil and corneal reflections were captured at 50 Hz by the eye tracker. The eye movement system was calibrated using a set of 9 screen locations so that gaze position could be calculated. The system produced videos of the stimuli with a mark at the gaze position for visual inspection. To avoid excessive head movements, the participants were sitting with their head placed on a chin rest. The videos of the participants' gaze position superimposed on the stimulus sequence were visually inspected and all trials in which participants' gaze deviated laterally more than approximately 1° visual angle from the fixation cross were excluded from further analysis. In no participants did this lead to exclusion of more than 15% of visual and audiovisual trials. Totally, 114 trials out of 1920 were excluded due to eye movements.

3. Results

The response percentages averaged across participants are displayed in Fig. 2. For both audiovisual and visual stimuli we analyze the effect of two factors, the attended visual articulation (V) and unilateral versus bilateral (UB) visual stimulus. We first transformed our data by taking the arcsine of the square root of the response ratios to homogenize the variances. We base our analysis on a two-way repeated measures ANOVA of transformed response ratios. We perform the analysis for each of the four response categories correcting the criterion of significance for multiple comparisons by $0.05/4 = 0.0125$ according to the Bonferroni method. Mauchly's test of violation of the sphericity assumption was not significant and the lower-bound method suggested no correction in the number of degrees of freedom.

3.1. Responses to audiovisual stimuli

For K-responses to the audiovisual stimuli, there was a significant main effect of factor V ($F(1,9) = 79.0$, $p < 10^{-5}$) which reflects that attending the face saying /aka/ increased the proportion of K-responses both in the unilateral (from 17% to 81%) and bilateral (from 32% to 63%) conditions. The V × UB interaction was significant ($F(1,9) = 16.1$, $p < 0.004$). This is because adding a distracting face saying /aka/ to an attended face saying /ata/ increased the proportion of K-responses whereas adding a distracting face saying /ata/ to an attended face saying /aka/ decreased the proportion of K-responses.

For P-responses to the audiovisual stimuli, no interaction or main effects were significant according to the Bonferroni corrected criterion of significance although both the V × UB interaction ($F(1,9) = 8.0$, $p > 0.02$) and the main effect of factor V ($F(1,9) = 7.9$, $p > 0.02$) were nearly significant.

For T-responses to the audiovisual stimuli, the main effect of factor V was significant ($F(1,9) = 27.6$, $p < 0.001$) reflecting that attending the face saying /ata/ increased the proportion of T-responses both in the unilateral (from 6% to 64%) and bilateral (from 16% to 33%) conditions. The significant V × UB interaction ($F(1,9) = 22.9$, $p < 0.001$) can be interpreted in the same way as for K-responses. Adding a distracting face saying /ata/ to an attended face saying /aka/ increased the proportion of T-

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

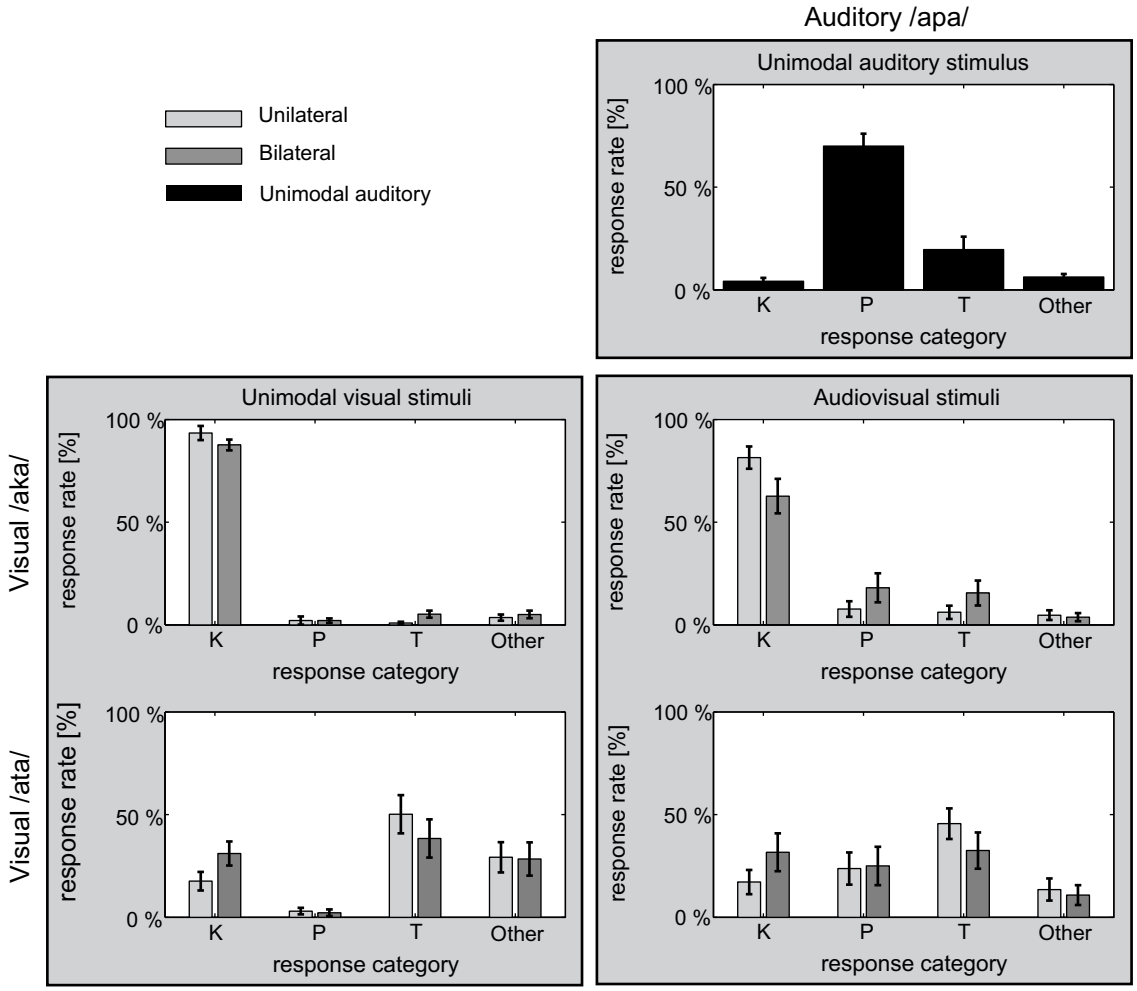

Fig. 2. Response proportions averaged across all participants. Error bars show the standard error of the mean. Plots organized so that rows determine the attended visual stimulus and columns determine the auditory stimulus. The top plot depicts results from unimodal auditory trials. Two leftmost plots depict results from unimodal visual trials. The remaining two plots depict results from audiovisual trials.

responses whereas adding a distracting face saying /aka/ to an attended face saying /ata/ decreased the proportion of T-responses.

For responses other than K, P and T, we found a significant main effect of factor V ($F(1,9) = 13.4,\ p < 0.006$) reflecting that the proportion of these responses was greater when attending the face saying /ata/ both in the unilateral (13% versus 5%) and bilateral (from 11% versus 4%) conditions.

3.2. Responses to visual stimuli

For K-responses to the visual stimuli, there was a significant main effect of factor V ($F(1,9) = 79.0,\ p < 10^{-4}$), which reflects that attending the face saying /aka/ increased the proportion of K-responses both in the unilateral (from 18% to 93%) and bilateral (from 31% to 88%) conditions. The V × UB interaction was significant ($F(1,9) = 16.1$, $p < 0.012$) and analogously to audiovisual stimuli, this is because adding a distracting face saying /aka/ to an attended face saying /ata/ increased the proportion of K-responses whereas adding a distracting face saying /ata/ to an attended face saying /aka/ decreased the proportion of K-responses.

For P-responses to the audiovisual stimuli, no main effects or interaction were significant ($p > 0.1$ for all effects) reflecting that the proportion of P-responses was very low at 3% or less for all visual conditions.

For T-responses to the visual stimuli, the significant main effect of factor V ($F(1,9) = 21.0,\ p < 0.002$) reflects that attending the face saying /ata/ increased the proportion of T-responses both in the unilateral (from 1% to 50%) and bilateral (from 5% to 38%) conditions. Again, the interpretation of the significant V × UB interaction ($F(1,9) = 10.6, p < 0.01$) is the same as described for audiovisual stimuli.

For responses other than K, P and T, we found a significant main effect of factor V ($F(1,9) = 13.4,\ p < 0.003$) reflecting that the proportion of these responses were greater when attending the face saying /ata/ for both in the unilateral (29% versus 4%) and bilateral (from 28% versus 5%) conditions.

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

4. Discussion

We found a strong effect of the object of visual attention in the critical audiovisual bilateral trials, in that participants gave more K-responses when attending the face saying /aka/ than when attending the face saying /ata/. This difference in K-responses was matched by a corresponding change in T-responses so that participants gave more T-responses when attending the face saying /ata/ than when attending the face saying /aka/. This shows that visual spatial attention could change the relative influence of the visual objects – i.e. the faces – on *auditory* perception. This exchange of influence between the visual objects could be due to an effect of attention already at the level of visual perception. This was confirmed by unimodal visual trials where participants lipread and also gave more K-responses when they attended the face saying /aka/ and more T-responses when they attended the face saying /ata/. This effect can be thought of as a visual analog to the cocktail party effect (Cherry, 1953), which is the ability to focus auditory attention on a single voice embedded in multi-speaker babble as we do at a busy cocktail party. The visual analog that we describe here is the ability to covertly focus visual attention on a single talking face in a crowd of talking faces. Furthermore, the proportion of P-responses did not depend on which face was attended. This means that the relative influence of audition and vision did not depend on the object of visual spatial attention. Therefore, in the current experiment, we ascribe the effect of the object of visual attention on audiovisual speech perception to an effect at the level of visual speech perception.

That the effect of attention on visual speech perception can propagate through audiovisual integration without influencing audiovisual integration is a novel finding. This dissociation between the object of visual attention and audiovisual integration was made possible by using two visual stimuli that both influenced the auditory speech percept. If one visual stimulus influences auditory speech much less than the other, then directing visual attention to it should decrease visual influence on the auditory speech percept. However, for unilateral audiovisual trials, participants gave more P-responses to visual /ata/ than to visual /aka/ indicating that visual /ata/ did not influence auditory perception as strongly as did visual /aka/. But, apparently, this effect was not strong enough to influence perception in bilateral trials. However, there was a non-significant tendency toward more P-responses in the bilateral condition compared to the unilateral condition when attending visual /aka/. If visual /ata/ was less strongly integrated with auditory speech, including it as a distractor face could have caused this tendency towards more P-responses as well as more T-responses.

We found a strong effect of visual attention on unimodal visual speech perception. Tiippana et al. found a similar effect but only for visual /t/ and not for visual /k/ and /p/. They suggested that this could be due to /t/ being difficult to discriminate from /k/ because it requires the detection of the tongue touching the alveolar ridge. When the tongue is not seen, the estimate of Finnish speakers is likely to default to /k/. Distinguishing /p/ from /k/ and /t/ is easy due to the prominent bilabial closure and might require less attention. This means that the effect of visual attention on unimodal visual speech perception was likely rather weak in Tiippana et al.'s study. Tiippana et al. noted a probable cause for this: In unimodal visual trials when attention was distracted, the participants were instructed to attend to the leaf but respond according to the face and thus had to report and attend to two different visual objects. This conflict might have caused participants to attend the face in unimodal visual trials even when instructed to direct attention towards the leaf. This could have weakened the effect of attention in unimodal visual trials. In the current study there was no such conflict in that participants never had to attend and respond to two different visual objects. This could explain why we found a stronger effect of attention in unimodal visual trials.

The current study distinguishes between the object and the load of attention. Notably, we found that the object of visual attention influences the auditory speech percept in the absence of any difference in cognitive and perceptual load. This extends the findings of Tiippana et al. who did not make the distinction between the object and the load of attention. Since Alsius et al. (2005) found that cognitive load can influence audiovisual integration there could have been an effect of the cognitive load of gaze tracking the leaf in Tiippana et al.'s study.

We did, however, find an effect of the perceptual load of the distractor face. This was seen when comparing responses to audiovisual bilateral and unilateral stimuli. We found that the attended face influenced auditory perception less in bilateral trials than in unilateral trials. This decrease in influence from the target face was matched by an increased influence from the distractor face. This suggests that attention occasionally or partly lapsed towards the distractor face when it was present. The influence of the voice showed a tendency to be greater in bilateral trials but this effect was non-significant. Our results are thus inconclusive on whether perceptual load influences audiovisual integration. Also, it should be noted that we did not vary perceptual load across a very large range. Greater variations in perceptual load might well cause greater and significant variations in the relative influence of audition and vision. The effect of perceptual load on audiovisual integration of speech thus remains a topic for future studies.

The role of cross-modal response bias is rarely discussed in studies of the McGurk illusion. In signal detection studies, a strong acoustic signal may induce a bias towards assuming that a visual signal also was present. This effect may occur in addition to a true perceptual effect – i.e. an acoustically induced change in visual sensitivity (Bolognini et al., 2005; Frassinetti et al., 2002). Likewise, the presence of a visual speech signal may bias observers' responses towards responding that the acoustic speech signal fell in

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

the same phonetic category as the visual speech signal in addition to any true perceptual McGurk effect. One argument for the true perceptual nature of the McGurk illusion is that the illusory percept may differ from both the auditory and visual percepts. Most studies have found that visual /k/ with auditory /p/ produce a McGurk illusion of hearing /t/. Such effects are called fusion effects. If the effect was that of a response bias, one might expect that perception be biased towards /k/ rather than /t/. However, the articulatory movements producing /k/ and /t/ are visually very similar in many talkers. So, if visual /k/ and /t/ are confused, visual /k/ might produce a response bias towards /t/ instead of /k/. So, apparent fusion effects might not guarantee that the effect is truly perceptual.

In this study, we used a variation of the McGurk illusion which we had found in a previous study (Andersen et al., 2001), where we found a stimulus set in which visual /k/ with auditory /p/ produced a McGurk illusion of hearing /k/. This is likely due to visual /k/ being very clear and distinguishable for this talker. Notably, this McGurk illusion occurred in an experiment where we also found the typical McGurk illusion of hearing /pt/ when presented with auditory /t/ and visual /p/. We include a sample bilateral audiovisual stimulus from the current experiment which contains this effect as online Supplementary material.

The results from the auditory only condition showed that observers could recognize the auditory /p/ reasonably well although the acoustic signal-to-noise ratio was rather low. This argues against observers relying solely on lip reading in the audiovisual trials when they were instructed to respond according to what they heard. However, auditory /p/ was confused with /t/ showing that it was not perfectly discriminable. In fact, it was less discriminable than auditory /k/ and /t/ which were almost always identified correctly. The reason for choosing such a low acoustic signal-to-noise ratio was to increase visual influence on auditory perception. This was necessary because the McGurk effect was, in fact, rather weak. In audiovisual bilateral trials, observers reported to hear /p/ veridically in 22% on average. This should be compared to the McGurk effect often completely capturing auditory perception with observers never reporting to have heard the acoustic speech signal veridically. Several factors might have contributed to this weakness of the McGurk effect: the eccentricity of the faces, slight asynchrony of the faces and of the voice and the complexity of the stimuli and task. Most likely it was due to a combination of these effects.

If we compare our findings with those of Soto-Faraco et al. (2004), we find an interesting asymmetry. They found that auditory selective attention could not alter visual influence on audition and concluded that vision influenced audition prior to auditory attentional selection. This agrees with the interpretation that since the McGurk illusion can elicit the MMN (Sams et al., 1991), auditory attention does not affect audiovisual interaction. These studies suggest that vision influences audition early in the auditory system, before auditory attention acts. This view is supported by studies showing that viewing articulatory movements can modify activity in primary auditory cortex (Ghazanfar et al., 2005; Pekkola et al., 2005) and even as early as in the brainstem (Musacchia et al., 2006). Contrary, we found that *visual* selective attention *did* alter visual influence on audition and conclude that vision influenced audition *after* visual attentional selection. This indicates that vision influences audition late in the visual system.

Visual spatial attention is likely to affect any cortical level that maintains spatial representations, so if integration occurs after spatial attention it is likely to occur at a high cortical level where the receptive fields are very wide. This would be after V1–V4 (Luck et al., 1997). This seems in concordance with lesions studies of audiovisual speech perception which found preserved visual influence on auditory speech even with extensive damage to V1–V4 (Campbell, 1992) and no visual influence on auditory speech perception with a lesion sparing V1–V4 (Campbell, 1996).

In a magnetoencephalographic study of brain activity in response to visual speech, Nishitani and Hari (2002) showed that early processing occur in occipital areas and then in the superior temporal sulcus (STS). The STS has been identified as an important brain area for audiovisual integration using neuroimaging techniques (Calvert, 2001; Calvert et al., 2000). It projects to areas containing auditory cortex in primates (Seltzer and Pandya, 1994). Furthermore, the laminar profile of visual inputs to auditory cortex in primates shows that they are feedback projections from a higher order cortex such as the STS (Schroeder and Foxe, 2002, 2005). Together, these results support a model of how visual speech influence auditory perception: Initial visual processing occurs in V1–V4 which projects to STS which, in turn, feeds back to the auditory pathway. This could happen in the auditory cortex or perhaps even earlier, in the brainstem. This model explains the asymmetry between the effects of auditory and visual attention. If visual attention modifies visual speech perception in V1–V4 then this effect would likely propagate through STS to the auditory pathway. This would explain our findings that visual attention can modify audiovisual as well as visual speech perception. If auditory attention modifies auditory speech processing after the influence from STS then it will operate on the visually modified auditory percept. This would explain why audiovisual integration of speech does not seem to be influenced by auditory attention.

In addition to phonetic content, auditory and visual speech contain information about emotional valence and the location of the speaker. There is ample evidence that audiovisual integration of these attributes occur and a number of studies have addressed the role of attention, arriving at the conclusion that this type of audiovisual integration seems to be pre-attentive.

Vroomen et al. (2001) studied the effect of performing an auxiliary task while judging the emotional valence of a spoken sentence combined with a static emotional facial expression. Whether the task was auditory or visual, the

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004

influence of the face on the perception of the emotional valence of the voice remained the same indicating that audiovisual integration was not influenced by selective attention.

Bertelson et al. studied the effect of endogenous attention on the ventriloquist effect, i.e. illusory displacement of a sound toward a concurrent visual stimulus. In one experiment they demonstrated the ventriloquist effect when a sound was perceived as originating from near the location of a visual stimulus displaced either to the right or to the left (Bertelson et al., 2000). They also imposed a perceptual task consisting of detecting a change in an auxiliary visual stimulus. The auxiliary visual stimulus was placed either at the location of the visual stimulus inducing the ventriloquist effect or displaced from it. Observers directed their gaze and thus attention to this auxiliary stimulus to conduct the task. Bertelson et al. found no effect of the location of the auxiliary stimulus on the ventriloquist effect indicating that audiovisual integration of spatial location does not depend on gaze direction and thus not on the direction of endogenous attention. In a follow-up study, Vroomen et al. studied the effect of exogenous attention on the ventriloquist effect (Vroomen et al., 2001). They noted that exogenous attention is a priori more likely to play a role in ventriloquism as the visual stimulus inducing the effect is likely to attract attention as well as perceptually bias the perceived location of the sound. However, Vroomen et al. found no effect of exogenous attention on the ventriloquist effect. Together these two studies provide strong evidence that the integration of spatial information from audiovisual stimuli is a pre-attentive process.

Talsma and Woldorff (2005) studied the effect of spatial attention on audiovisual integration using electroencephalography and non-speech stimuli. They presented auditory, visual or audiovisual stimuli either to the left or right visual hemifield with observers attending to one hemifield. They compared the event-related potentials (ERP) elicited by audiovisual stimuli to the sum of the ERPs elicited by auditory and visual stimuli separately. They found a significant difference, which reflects audiovisual interaction effects. Notably, this effect was greater for the attended hemifield indicating that the audiovisual interaction effects were modified by attention. Since Talsma and Woldorff's paradigm is similar to ours and their conclusion opposite, we find it interesting to speculate on the cause of this difference. In Talsma and Woldorff's study, attention was either directed towards the stimulus or away from it. This is likely to affect perception of unimodal auditory and visual stimuli as well as perception of audiovisual stimuli. In accordance, Talsma and Woldorff also reported effects of attention on the ERPs from unimodal auditory and visual stimuli. This could influence the relative reliability of the unimodal percepts, which can have a strong influence on audiovisual integration (Andersen et al., 2004; Warren, 1979). Therefore, it might only be possible to dissociate attention and audiovisual integration if attention does not strongly influence the reliability of the unimodal auditory and visual stimuli. In our study, spatial attention should not influence the reliability of the auditory stimulus, which was always presented at the same distance from the focus of attention, but attention did change the reliability of the visual stimulus somewhat, because visual /ata/ was not as reliable as visual /aka/. However, this effect was apparently not strong enough to change audiovisual integration. Another possible reason for the differences between the two studies is that different mechanisms may govern attention and audiovisual integration for speech and non-speech stimuli.

In summary, our main results show that visual spatial attention can have a strong influence on visual as well as audiovisual speech perception without influencing audiovisual integration. This suggests that the effect of visual spatial attention occurs at the level of visual perception and propagates trough audiovisual integration to influence auditory perception. To use two popular idioms of cognitive science: The visual cocktail party effect propagates to the McGurk illusion.

Acknowledgements

The authors thank Mr. Mikko Viinikainen and Ms. Jaana Simola for practical assistance. The study was financially supported by the Academy of Finland, the MUHCI European Union Research Training Network and the Danish Council for Strategic Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.specom.2008.07.004.

References

Alsius, A., Navarra, J., Campbell, R., Soto-Faraco, S., 2005. Audiovisual integration of speech falters under high attention demands. Curr. Biol. 15 (9), 839–843.

Andersen, T.S., Tiippana, K., Lampinen, J., Sams, M., 2001. Modelling of audiovisual speech perception in noise. In: Proc. AVSP 2001, pp. 172–176.

Andersen, T.S., Tiippana, K., Sams, M., 2002. Using the fuzzy logical model of perception in measuring integration of audiovisual speech in humans. In: Proc. NF2002.

Andersen, T.S., Tiippana, K., Sams, M., 2004. Factors influencing audiovisual fission and fusion illusions. Cogn. Brain Res. 21 (3), 301–308.

Bertelson, P., Vroomen, J., de Gelder, B., Driver, J., 2000. The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. 62 (2), 321–332.

Bolognini, N., Frassinetti, F., Serino, A., Ladavas, E., 2005. "Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs. Exp. Brain Res. 160 (3), 273–282.

Calvert, G.A., 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex 11 (12), 1110–1123.

Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. 10 (11), 649–657.

Campbell, R., 1992. The neuropsychology of lipreading. Philos. Trans. Roy. Soc. London B Biological Sci. 335 (1273), 39–44, discussion 44-35.

Campbell, R., 1996. Seeing brains reading speech: a review and speculations. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines: Models. In: Systems and Applications. Springer, Berlin, pp. 115–134.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Amer. 25, 975–979.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., Deltenre, P., 2002. Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin. Neurophysiol. 113 (4), 495–506.

Crowther, C.S., Batchelder, W.H., Hu, X., 1995. A measurement-theoretic analysis of the fuzzy logic model of perception. Psychol. Rev. 102 (2), 396–408.

Cutting, J.E., Bruno, N., Brady, N.P., Moore, C., 1992. Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth. J. Exp. Psychol. Gen. 121 (3), 364–381.

Frassinetti, F., Bolognini, N., Ladavas, E., 2002. Enhancement of visual perception by crossmodal visuo-auditory interaction. Exp. Brain Res. 147 (3), 332–343.

Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K., 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J. Neurosci. 25 (20), 5004–5012.

Grant, K.W., Seitz, P.F., 1998. Measures of auditory-visual integration in nonsense syllables and sentences. J. Acoust. Soc. Amer. 104 (4), 2438–2450.

Lavie, N., 2005. Distracted and confused: selective attention under load. Trends Cogn. Sci. 9 (2), 75–82.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36.

Liberman, A.M., Mattingly, I.G., 1989. A specialization for speech perception. Science 243 (4890), 489–494.

Luck, S.J., Chelazzi, L., Hillyard, S.A., Desimone, R., 1997. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol. 77 (1), 24–42.

Massaro, D.W., 1984. Children's perception of visual and auditory speech. Child Develop. 55, 1777–1788.

Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum, Hillsdale, NJ.

Massaro, D.W., 1998. Perceiving Talking Faces. MIT Press, Cambridge, Massachusetts.

Massaro, D.W., Cohen, M.M., Smeele, P.M., 1996. Perception of asynchronous and conflicting visual and auditory speech. J. Acoust. Soc. Amer. 100 (3), 1777–1786.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Möttönen, R., Krause, C.M., Tiippana, K., Sams, M., 2002. Processing of changes in visual speech in the human auditory cortex. Brain Res. Cogn. Brain Res. 13 (3), 417–425.

Munhall, K.G., Gribble, P., Sacco, L., Ward, M., 1996. Temporal constraints on the McGurk effect. Percept. Psychophys. 58 (3), 351–362.

Musacchia, G., Sams, M., Nicol, T., Kraus, N., 2006. Seeing speech affects acoustic information processing in the human brainstem. Exp. Brain Res. 168 (1–2), 1–10.

Näätänen, R., 1992. Attention and Brain Function. Lawrence Erlbaum Associates, Hillsdale, NJ.

Nishitani, N., Hari, R., 2002. Viewing lip forms: cortical dynamics. Neuron 36 (6), 1211–1220.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I., Möttönen, R., Tarkiainen, A., et al., 2005. Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport 16 (2), 125–128.

Pitt, M.A., 1995. Data fitting and detection theory: reply to Massaro and Oden (1995). J. Exp. Psychol. Learn. Mem. Cogn. 21 (4), 1065–1067.

Posner, M.I., 1980. Orienting of attention. Quart. J. Exp. Psychol. 32 (1), 3–25.

Posner, M.I., Snyder, C.R., Davidson, B.J., 1980. Attention and the detection of signals. J. Exp. Psychol. 109 (2), 160–174.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. Science 212 (4497), 947–949.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., et al., 1991. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci. Lett. 127 (1), 141–145.

Schroeder, C.E., Foxe, J.J., 2002. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Brain Res. Cogn. Brain Res. 14 (1), 187–198.

Schroeder, C.E., Foxe, J., 2005. Multisensory contributions to low-level, 'unisensory' processing. Curr. Opin. Neurobiol. 15 (4), 454–458.

Schwartz, J.L., 2006. The 0/0 problem in the fuzzy-logical model of perception. J. Acoust. Soc. Amer. 120 (4), 1795–1798.

Seltzer, B., Pandya, D.N., 1994. Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J. Comp. Neurol. 343 (3), 445–463.

Soto-Faraco, S., Navarra, J., Alsius, A., 2004. Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. Cognition 92 (3), B13–B23.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Amer. 26 (2), 212–215.

Talsma, D., Woldorff, M.G., 2005. Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. J. Cogn. Neurosci. 17 (7), 1098–1114.

Tiippana, K., Andersen, T.S., Sams, M., 2004. Visual attention modulates audiovisual speech perception. Eur. J. Cogn. Psychol. 16 (3), 457–472.

Treisman, A., Gelade, G., 1980. A feature integration theory of attention. Cogn. Psychol. 12 (1), 97–136.

Tuomainen, J., Andersen, T.S., Tiippana, K., Sams, M., 2005. Audio-visual speech perception is special. Cognition 96 (1), B13–B22.

Vroomen, J., de Gelder, B., 2000. Crossmodal integration: a good fit is no criterion. Trends Cogn. Sci. 4 (2), 37–38.

Vroomen, J., Bertelson, P., de Gelder, B., 2001a. The ventriloquist effect does not depend on the direction of automatic visual attention. Percept. Psychophys. 63 (4), 651–659.

Vroomen, J., Driver, J., de Gelder, B., 2001b. Is cross-modal integration of emotional expressions independent of attentional resources? Cogn. Affect Behav. Neurosci. 1 (4), 382–387.

Warren, D.H., 1979. Spatial localization under conflict conditions: is there a single explanation? Perception 8 (3), 323–337.

Please cite this article in press as: Andersen, T.S. et al., The role of visual spatial attention in audiovisual speech perception, Speech Comm. (2008), doi:10.1016/j.specom.2008.07.004